Multimodal Bayesian neural networks

Weijiang Xiong, Markus Heinonen, Samuel Kaski

June 1, 2021

This is the running document of the summer intern project 2021 of Weijiang Xiong on Bayesian neural networks. The work is supervised by PhD Markus Heinonen under Prof. Samuel Kaski and with help of Msc Trinh Trung.

TLDR: Develop multimodal inference for BNNs.

1 Research diary

TODO for Weijiang

2 Problem definition

2.1 BNN setting

Bayesian neural networks are a probabilistic variants of neural networks, where conventionally we place a prior $p(\mathbf{w})$ on the neural weights $\mathbf{w}$, which results in a posterior distribution $p(\mathbf{w}|\mathcal{D})$ of the weights given data $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$. We assume that the weights $\mathbf{w}$ contains all parameters of the network, including biases, filter kernels, etc.

While conventional DNNs only account for a single function hypotheses $f(\mathbf{x}; \mathbf{w}_{\text{opt}})$, BNNs can capture a large distribution of alternative function f hypotheses by considering multiple convolution, weight and pattern combinations. This is described by the predictive posterior

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) = \int \underbrace{p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{w})}_{\text{likelihood}} p(\mathbf{w}|\mathcal{D}) d\mathbf{w}, \tag{1}$$

where we average predictions for a new test point $(\mathbf{x}_*, \mathbf{y}_*)$ from all posterior weights. This generally has been shown to improve calibration, uncertainty quantification, out-of-distribution prediction and robustness. For excellent overviews of BNNs, we refer to Wilson's work (Wilson, 2020; Wilson and Izmailov, 2020). The loss landscapes have been studied by Garipov et al. (2018)

Most of BNN research has focused on five major research questions

- What type of priors should be used?
- How to do model selection (that is, learn hyperparameters)
- How to numerically infer the posterior?
- How should we parameterise the neural function?
- What are theoretical connections of probabilistic neural networks to other models?

In this summer project we focus on the inference problem (and also a bit on the parameterisation).

2.2 Variational inference for BNNs (BNN-VI)

The currently dominant BNN approach uses Gaussian priors $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma^2 I)$ for the weights, and applies mean-field variational inference (MFVI) (Blundell et al., 2015). In MFVI full posterior inference is sidestepped by introducing a tractable variational posterior approximation $q(\mathbf{w})$ such that its Kullback Leibler distance to the true posterior is minimized,

$$\arg\min_{\mathbf{w}} \mathrm{KL}[q(\mathbf{w})||p(\mathbf{w}|\mathcal{D})]. \tag{2}$$

Under variational inference (See Blei et al. (2017)) it can be shown that this intractable KL is (remarkably!) equivalent to maximizing the evidence lower bound

$$\log p(\mathcal{D}) \geq \underbrace{\mathbb{E}_{q(\mathbf{w})} \log p(\mathcal{D}|\mathbf{w})}_{\text{variational likelihood}} - \underbrace{\mathrm{KL}[q(\mathbf{w})||p(\mathbf{w})]}_{\text{KL term}}, \tag{3}$$

where we now maximize the variational likelihood and minimize the KL term between the posterior approximation $q(\mathbf{w})$ and the weight prior $p(\mathbf{w})$. By assuming factorised Gaussians

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \operatorname{diag} \mathbf{s}^2) \tag{4}$$

$$= \prod_i \mathcal{N}(w_i|m_i, s_i^2) \tag{5}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \sigma^2 I) \tag{6}$$

$$= \prod_i \mathcal{N}(w_i|0, \sigma^2) \tag{7}$$

both the KL term and the variational likelihood term become easy to compute and optimise (via the reparameterisation trick).

2.3 Alternative inference methods

While the MFVI-BNN is the 'standard' approach, alternative approaches have also been proposed. Full-batch HMC samples directly from the extremely complex posterior landscape $p(\mathbf{w}|\mathcal{D})$ (Izmailov et al., 2021), which is highly impractical. Earlier Wenzel et al. (2020) proposed more practical minibatch-based SG-HMC to infer the same posterior, and found out the need to apply tempering to downplay the prior. The celebrated method SWAG uses the SGD optimisation trace to estimate the shape of the local optimum the SGD is converging towards, and places a Gaussian approximation on it (This method comes standard in pytorch 1.6[1]) (Maddox et al., 2019). MultiSWAG repeats the procedure for multiple initialisations following the deep ensembles. Early works have also used normalizing flows to estimate possibly multimodal posteriors (Louizos and Welling, 2017).

2.4 Low-rank parameterisations

The inference of the posterior $p(\mathbf{w}|\mathcal{D})$ is inherently extremely difficult to the weights being often up to 100-million dimensional vectors. Exploration of the loss landscape in such high-dimensional spaces is possibly a futile task. A series of works have sought low-rank parameterisations of neural networks (Dusenberry et al., 2020; Karaletsos et al., 2018; Karaletsos and Bui, 2020; Trung et al., 2021). Multiple authors have proposed learning multiplicative latent *node* variables,

$$\mathbf{x}_{\ell+1} = \sigma(W_\ell(\mathbf{x}_\ell \circ \mathbf{z}_\ell) + \mathbf{b}_\ell), \tag{8}$$

where the node activations $\mathbf{x}_\ell$ of layer ℓ are multiplied by stochastic random variables $\mathbf{z}$. By defining the weights are deterministic parameters, we only need to infer the posterior $p(\mathbf{z}|\mathcal{D})$ of a much smaller latent variable space.

[1] https://pytorch.org/blog/pytorch-1.6-now-includes-stochastic-weight-averaging/

2.5 Problem definition

The topic of the summer project is to develop multimodal posterior inference for BNNs in combination with a node parameterisation.

Multimodal inference. We need to select a way to perform multimodal inference. We invented a new mixture-of-Gaussian variational approximation in our iBNN paper (Trung et al., 2021), which however does not explore many modes. Conventional approximate inference literature would propose the boosting VI approach, or even active learning. The normalizing flows are a great candidate for multimodality, however there are numerous NF models to choose from (Papamakarios et al., 2021).

Node parameterisations. We can start from the iBNN code and its node parameters, but alternatives also exist (such as rank-1 BNN (Dusenberry et al., 2020)).

Success criteria. We want to compare our method against BNN literature on MNIST, Fashion-MNIST and CIFAR problems. We want to showcase the improvements of the multimodal inference in accuracy (error), calibration (ECE, log-likelihood), robustness (corruption) and out-of-distribution performance. Our main method to beat deep ensembles, and we want to show how the degree of multimodality correlates with better results.

2.6 How to start

At the start of the project

- Let's invite and setup PML slack for the project, and regular project meetings
- Test and play around with the iBNN codebase and Aalto's computing clusters. Familiarize yourself with the code and repeat some of the experiments of the iBNN as a demonstration
- Read iBNN paper (Trung et al., 2021) and Wilson's two generalisation papers (Izmailov et al., 2021; Wilson and Izmailov, 2020)
- Try SGHMC as baseline approach: how does it work? Use an existing implementation to do this (eg. Wenzel et al. (2020)).
- Implement in pytorch the multimodal inference (eg. normalizing flow)

References

David Blei, Alp Kucukelbir, and Jon McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *ICML*, 2015.

Michael W. Dusenberry, Ghassen Jerfel, Yeming Wen, Yi-An Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable Bayesian neural nets with rank-1 factors. In *ICML*, 2020.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *NIPS*, 2018.

Pavel Izmailov, Sharad Vikram, Matthew Hoffman, and Andrew Wilson. What are bayesian neural network posteriors really like? *arxiv*, 2021.

Theofanis Karaletsos and Thang D. Bui. Hierarchical Gaussian process priors for Bayesian neural network weights. In *NIPS AABI Workshop*, 2020.

Theofanis Karaletsos, Peter Dayan, and Zoubin Ghahramani. Probabilistic meta-representations of neural networks. *arXiv*, 2018.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In *ICML*, 2017.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *NIPS*, 2019.

George Papamakarios, Eric Nalisnick, Danilo Rezende, Shakir Mohamed, and Balaji Lakshminarayana. Normalizing flows for probabilistic modeling and inference. *JMLR*, 2021.

Trinh Trung, Markus Heinonen, and Samuel Kaski. Scalable bayesian neural networks by layer-wise input augmentation. *arxiv*, 2021.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In *ICML*, 2020.

Andrew Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *NIPS*, 2020.

Andrew Gordon Wilson. The case for Bayesian deep learning. *arXiv:2001.10995*, 2020.